URBAN ACOUSTIC SCENE CLASSIFICATION

Priya Khokher, Adhlere Coffy, Maria Filippelli, Manushi Majumdar, Neil D'Souza

ABSTRACT

Sound is an essential sense that is often undervalued especially in urban analysis but it is a very unique signature of a city. There is some research done for detection of acoustic events but very little done in terms of sound scenes which give a more detailed picture of the city. In this paper we explore the methods to extract features from sound scenes and compare different machine learning algorithms to find the one best suited for the task of classifying urban acoustic scenes. Sound data from two different cities is also used to see how unique the sound signature of a city is. The paper shows that Support Vector Machines are ideal for classifying sound scenes as well as the need for more research in the area of urban acoustics but fails to truly confirm science of cities in terms of sound..

1. INTRODUCTION

Cities are filled with an abundance of sensory perceptions and, at times, the cacophony of sounds can prove to be overwhelming. But can this myriad of noises be utilized to identify a particular geographical area or "urban scene"? Can cities be distinguished from each other based on the sound scenes that are prevalent within them? This undertaking sought to execute preliminary Machine Learning analyses to investigate the question: Are urban sound profiles from different cities similar enough to accurately classify, regardless of origin?

To examine this question, this team classified common city sounds utilizing a variety of machine learning models in an effort to determine which achieves the greatest accuracy. Sound scenes were taken from datasets developed in London and Paris and transformed using Mel-frequency cepstral coefficients (MFCCs) in order to create values to test with Naive Bayes, Support Vector Machines, and Neural Networking models. These models were chosen because they are the most applicable to sound scene classification since they are better at handling the kind of features that are involved with sound (i.e specific numeric MFCC values). They have also been overshadowed by Random Forest methods in the recent past and this aims to see which of the older ones performs the best.

From the datasets five city scenes common to both datasets were used for classification, a subway station, restaurant, busy street, farmer's/open air market, and a quiet street. An additional group was reserved for "general" scenes that were not cross compatible between datasets. These scenes were selected for the wide range of urban activities that they represented, are distinctly different from each other, and presence in both datasets.

The next section goes over the prior research similar to this project followed by explanation of the datasets used. Extraction of features from sound is explained before discussing the models. Finally the results and presented and discussed.

2. LITERATURE REVIEW

In the paper 'A dataset and Taxonomy for Urban Sound Research' by Justin Salamon et al they addressed the lack of a common taxonomy and the scarceness of large, real-world, annotated sound data. They presented the Urban Sound Taxonomy and UrbanSound (dataset). Through a series of classification experiments the team studied the challenges presented by the dataset, and identified avenues for future research.[1]

Another paper titled 'Unsupervised Feature Learning for Urban Sound Classification' by Justin Salamon and Juan Pablo Bello discusses a study of the application of unsupervised feature learning to urban sound classification. They apply the spherical k-means algorithm on the public dataset available for sound and compare it to a baseline system based on MFCCs.[2] The same team published another paper titled 'Feature Learning with Deep Scattering for Urban

Sound Analysis', wherein they evaluated the use of the scattering transform as an alternative to the mel spectrogram in the context of unsupervised feature learning for environmental sound classification. They demonstrated that comparable (or slightly better) performance can be obtained using the scattering transform whilst reducing both the amount of training data required for feature learning and the size of the learned codebook by an order of magnitude.[3]

The existing research focuses on classifying sound sources. While sound sources merely focus on extracting the information of the source of that sound, through our project we intend to classify sound scenes, which include background sounds present in the sound clips. The classification of sound scenes could open doors to deeper research in this field.

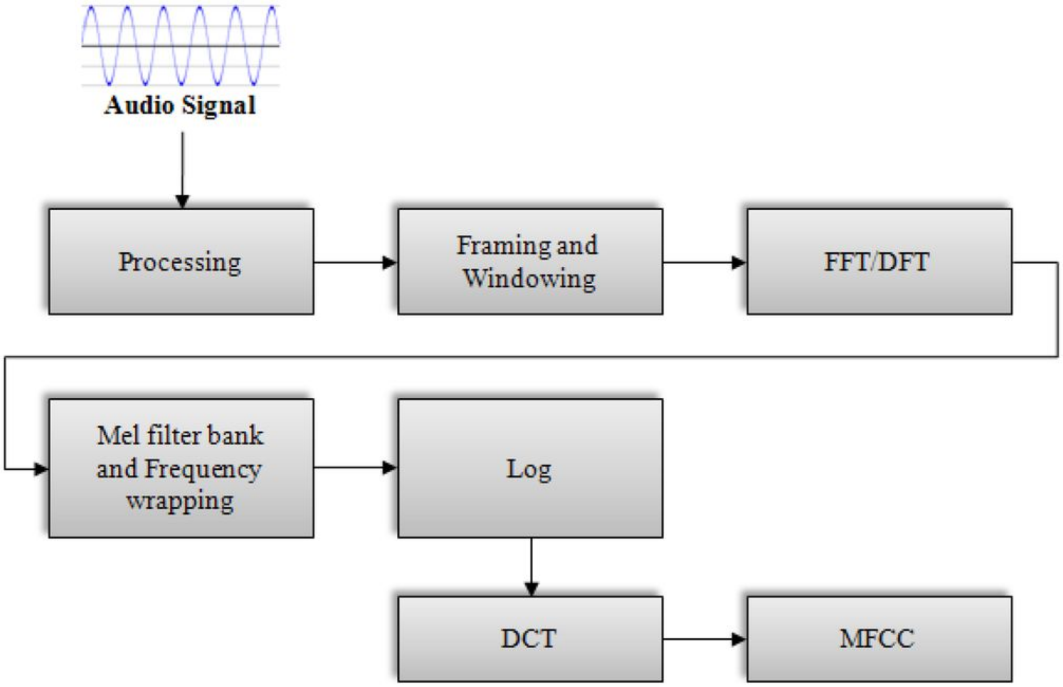
3. METHOD

3.1. Datasets

The two datasets used are from London and Paris sound scene classifications. The Detection and Classification of Acoustic Scenes and Events (DCASE) Queen Mary dataset, a sound scene classification experiment done in London, comprises of 10 scene classes with 20 sound clips each. The Rouen dataset is from a sound scene classification experiment in Paris and has 19 sound scenes. Several sound scenes exist in both datasets making them ideal to train, validate, and test with our models. The scenes common to both and chosen for classification are bus, office, park, supermarket and tube station; all other scenes are classified as other. The sound clips from both datasets are 30 seconds each which is a minimum length for testing.

3.2. Feature Extraction

Classification of urban sound scenes entailed categorization of data points based on their predefined set of values. Sounds scenes are transformed into a feature space of frequency and amplitude and required several transformations to prepare the data for model testing. Creating a feature space for sound dataset, in the frequency domain is better than amplitude as frequencies are more specific and inherent to a sound type than amplitude which can be externally manipulated. So for instance, DFT could be natural choice. As sound datasets can contain many redundancies in information, it is necessary to choose a feature set as invariant to changes within the natural class' range as possible. Therefore the next step is to produce Mel Frequency Cepstral Coefficients (MFCCs) which is a best practice for feature extraction when using audio data. These MFCCs have proven to form a more robust feature space and has been used for the project too. MFCCs were calculated in three steps, first a fourier transformation was performed, then powers were mapped onto a Mel scale, and finally the MFCC of each spectrum was produced.



The fourier transformation divided each sound scene into 20 millisecond frames. This was important because sound scenes differ in the patterns they produce and no reasonable computation can be performed by taking the mean of longer frames. The fourier transform is then used to compute the frequency spectrum of the frame.

Mel scaling related the perceived frequency of a pure tone to its actual measured frequency. Incorporating this scale makes the features match more closely what humans hear. To achieve this task, 40 different mel filter banks were used to bin each frequency change, ranging from fmin = 0 to fmax = 22050 Hz.

To find the resulting MFCCs, the log of each Mel power frequency was taken and then the discrete cosine transformation was taken. For each set of MFCCs, each frame (which were 40, the last 20 were kept as these capture less noise), a summarization of : minimum, maximum, median, mean, variance, and kurtosis. From this the mean and variance of the first derivatives were chosen as features.

3.3. Machine Learning

After extracting the MFCC and calculating appropriate summary statistics (e.g. mean, median, variance, minimum, maximum, standard deviation, kurtosis), the data was separated into two main segments. In the first segment, the training and validation sets were exclusively comprised of the sound data from Paris while the test set was compiled from London data. The resulting segment (training, validation and testing) was used to evaluate whether or not the Science of Cities (SOC) hypothesis can be adapted for sound and used as an alternate approach to understanding urban environments.

In the second segment, the data was randomly shuffled such that the training, validation, and test sets reflected all of the available data. This was achieved by first, concatenating the train and validation set before then randomly splitting the resulting dataset such that the validation data varied in proportional representation from 20 - 30 percent of the full dataset.

This two-prong approach of testing the generalizability of the models is ideal for testing the coverage of classifications parameters and the ability to “respond” to randomly organized data constituting multiple urban sound sources.

3.3.1. Naive Bayes

Naive Bayes (NB) was performed first because of its streamlined implementation and classification methodology. For this task, a continuous NB classifier was used to account for the number of scenes identified for classification and the corresponding feature set which was quite expansive. A primary consequence to take note of when using NB is that it assumes that features are independent of one another. This failure to identify correlations between features lends NB to underperform compared to other machine learning techniques, demonstrated in this project.

Using the aforementioned dataset split, the accuracy of the classifier of the validation and test sets was compared. The first dataset validation/test accuracy was an underwhelming 68%, a less than stellar performance. The second dataset yielded an accuracy of 22% thus proving that Naive Bayes was not a good classifier for the urban sound scenes.

3.3.2. Neural Networks

Neural Networks (NN) provide an intuitive, extremely flexible option for classifying data and making it an actionable all in one process. The difficulty that arises from the use of NN’s is the time consuming nature of optimizing the performance. NN’s are very easy to implement and use but quite difficult to improve for the sake of finding the “best” outcome. Adjusting parameters such as learning rate, number of hidden layers, and batch size can be very time consuming without yielding results that demonstrate improvements in performance over time. Being based on the logistic regression, NN inherently utilizes probabilistic relationships between inputs, unlike NB, and does not assume that all inputs are independent of each other. A minor adjustment to the negative log likelihood function of utilizing the mean as opposed to the sum allows for reduction in dependence of the learning rate on the batch size. For this application, it was of interest to identify how well the NN model could learn from given sound scenes with a wide range of input magnitudes without severely impacting computational time and/or resources. Finale parameters for the model were as follows: Learning Rate - 0.85, Batch Size - 400, and 10 hidden layers.

3.3.3. Support Vector Machines

Support Vector Machines (SVM) uses a technique called the kernel trick to transform the data and then, based on these transformations, compute an optimal boundary between the possible outputs discovered. SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.The parameters tweaked in the model are explained as follows.

C represents soft margin constant. Softness in margins (a low value) allows for errors to be made while fitting the model to the training data set. Conversely, hard margins will result in fitting of a model that allows zero errors. Sometimes it can be helpful to allow for errors in the training set, because it may produce a more generalizable model when applied to new datasets. Forcing rigid margins though can result in a model that performs perfectly in the training set, but is possibly over-fit/less generalizable when applied to a new dataset. C acts as a penalty parameter. If it is too large, we have a high penalty for non separable points and we may store many support vectors and overfit. If it is too small, we may have underfitting.

The linear and rbf (gaussian) kernel SVM models had a lower accuracy than the polynomial kernel model. Higher degree polynomial kernels allow a more flexible decision boundary, although in this case it could suggest overfitting of data. Hence we consider polynomial degree to be 2, making it quadratic. Changing the degree from 2 to 3, increases the accuracy of the prediction using poly kernel by 10%. Selecting a degree value of 3 allows for a more flexible decision boundary without overfitting the data. The best accuracy is observed for C = 2.5 and when the training data is split approximately in 75:25 ratio.

4. RESULTS

MODELS (Best Split)	**ACCURACY**	
	SOC Set	Generalizability Set
Gaussian Naive Bayes (50:50)	68.4 %	21.8 %
Neural Networks (70:30)	78.0 %	80.0 %
Polynomial Support Vector Machine (~75:25)	93.7 %	91. 4%

The model that performs best on the validation set in both cases is SVM with a 3 degree polynomial kernel and penalty factor of 2.5. This is the model that is used on the test set as shown below.

TESTING	37.5 %	71.2%

Thus we can say that sound may not be the best way to demonstrate science of cities though it still shows some support. The accuracy over the generalized data is also very promising.

5. CONCLUSIONS

The SONYC project at CUSP has already delved into identifying sound sources in NYC, the next step would be to understand and identify the sound scenes which will provide the context for different sources. The classification will provide a basis for further research in the field of urban acoustics/sounds, which is still being explored. The primary take away, and relative success, of the Support Vector Machine model in the project execution is the implication that cities are in fact relatable in terms of typical sound scenes. The number of assumptions made here can be easily addressed with minimal effort made on the part of the experimenters, those being; increase of sample size and rate, additional sound profiles datasets from comparably sized cities, greater duration of time spent refining the models to ensure optimization has been met, and perhaps more extensive actions taken to randomize the information such that the degree of generalizability can be maximized.

6. REFERENCES

Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.

Salamon, Justin, and Juan Pablo Bello. "Feature learning with deep scattering for urban sound analysis." *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015.

Salamon, Justin, and Juan Pablo Bello. "Unsupervised feature learning for urban sound classification." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

Datasets

A. Rakotomamonjy, G. Gasso, Histogram of gradients of Time-Frequency representations for audio scene detection, Technical report, HAL, 2014

D. Giannoulis, E. Benetos, D. Stowell, and M. D. Plumbley, "IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Public Dataset for Scene Classification Task", Queen Mary University of London, 2012.

Code

https://github.com/neilverosh/UASSC